WHAT EVOLUTION IS 2

THIS CHAPTER introduces three basic building blocks of evolutionary dynamics: replication, selection, and mutation. These are the fundamental and defining principles of biological systems. They apply to any biological organization anywhere in our or other universes and do not depend on the particular details of which chemistry was recruited to embody life. Any living organism has arisen and is continually modified by these three principles.

Evolution requires populations of reproducing individuals. In the right environment, biological entities, such as viruses, cells, and multicellular organisms can make copies of themselves. The blueprint that determines their structure, the genomic material in form of DNA or RNA, is replicated and passed on to the offspring. Selection results when different types of individuals compete with each other. One type may reproduce faster and thereby outcompete the others. Reproduction is not perfect, but involves occasional mistakes, or mutations. Mutation is responsible for generating different types that can be evaluated in the selection process, and thus results in biological novelty and diversity. Selection will choose to maintain some innovations and dismiss others, and can favor or oppose genetic diversity.

At the end of this chapter we will focus on the Hardy-Weinberg law of random mating. This discussion will be our only venture into the mathematics of sexual reproduction. In subsequent chapters we will encounter additional principles of evolutionary dynamics, such as random drift and spatial movement.

2.1 REPRODUCTION

Imagine a single bacterial cell in a perfect environment that contains all the nutrients required for growth and happiness. In this bacterial heaven, the fortunate cell and all its offspring divide every 20 minutes, which is the known world record for bacterial cell division in an ideal lab setting. After 20 minutes the cell has given rise to 2 daughter cells. After 40 minutes there are 4 granddaughters, and after one hour there are 8 great granddaughters. How many cells will there be after three days?

After t generations there are 2^t cells. In three days there are 216 generations. Hence we expect $2^{216} = 10^{65}$ cells. The total mass of these cells would exceed the mass of the earth by many orders of magnitude.

The growth law for this overwhelming expansion can be written as a recursive equation

$$x_{t+1} = 2x_t. \tag{2.1}$$

Here x_t is the number of cells at time t, and x_{t+1} is the number of cells at time $t + 1$. The equation means that at time $t + 1$ there are twice as many cells as at time t. Time is measured in numbers of generations.

The number of cells at time 0 is given by x_0. With this initial condition, the solution of equation (2.1) can be written as

$$x_t = x_0 2^t. \tag{2.2}$$

Equation (2.1) is a so-called difference equation, because time is measured in discrete steps.

We can also formulate a differential equation for exponential growth that measures time as a continuous quantity. Let $x(t)$ denote the abundance of cells at time t. Suppose that cells divide at rate r. More precisely, we assume that the

time for cell division follows an exponential distribution with average $1/r$. We can write the differential equation

$$\dot{x} = \frac{dx}{dt} = rx. \tag{2.3}$$

Throughout this book, I will use the standard notation $\dot{x}$ to refer to differentiation (of x) with respect to time. If the abundance of cells at time 0 is given by x_0 then the solution of the differential equation (2.3) is

$$x(t) = x_0 e^{rt}. \tag{2.4}$$

Let us reconsider our bacterial supernova. If we measure time in units of days, then $r = 72$ means that the time for a cell cycle requires, on average, 20 minutes (calculated by dividing the total number of minutes in a day, 1,440, by 72). Hence there are 72 cell divisions in one day. After three days, one bacterial cell has generated e^{216} cells which is approximately 6×10^{93} cells.

The discrepancy between the differential equation and the difference equation is a consequence of the varying assumptions for the distribution of the generation time. The difference equation assumes that each cell division occurs after exactly 20 minutes. The differential equation assumes that each cell division occurs after a time which is exponentially distributed around an average of 20 minutes. The exponential distribution is defined as follows: the probability that cell division occurs between time 0 and τ is given by $1 - e^{-r\tau}$. On average, cells divide after $1/r$ time units.

So far we have ignored cell death. Let us now suppose that cells die at rate d, which means that they have an exponentially distributed lifespan with an average of $1/d$. The differential equation becomes

$$\dot{x} = (r - d)x. \tag{2.5}$$

The effective growth rate is the difference between the birth rate, r, and the death rate, d. If $r > d$, then the population will expand indefinitely. If $r < d$, then the population will converge to zero and become extinct. If $r = d$, then the population size remains constant, but this situation is unstable: small deviations from absolute equality between birth and death will lead to either exponential expansion or decline. It is important to note that setting $r = d$

in equation (2.5) does not constitute a mechanism for maintaining a stable constant population size.

The simple equation (2.5) allows us to introduce an extremely important concept in evolution, ecology and epidemiology: the basic reproductive ratio, r/d. This ratio denotes the expected number of offspring that come from any one individual. The average lifetime of a cell is $1/d$. The rate of producing offspring cells is given by r. If each cell produces on average more than one offspring, $r/d > 1$, then an exponential expansion will follow. A basic reproductive ratio greater than one is a necessary condition for population expansion.

We have observed that ongoing exponential growth can lead to unreasonably high numbers in a very short time. In a realistic environment, the expanding population will hit constraints that prevent further expansion. For example, the population might run out of nutrients or physical space.

A model for population expansion with a maximum carrying capacity is given by the logistic equation

$$\dot{x} = rx(1 - x/K). \tag{2.6}$$

As before, the parameter r refers to the rate of reproduction in the absence of density regulation, when the population size, x, is much smaller than the carrying capacity K. As x increases, the rate of growth slows down. When x reaches the carrying capacity, K, then the population expansion ceases. For the initial condition x_0, the solution of equation (2.6) is given by

$$x(t) = \frac{Kx_0e^{rt}}{K + x_0(e^{rt} - 1)}. \tag{2.7}$$

In the limit of infinite time, $t \to \infty$, the population size converges to the equilibrium $x^* = K$. Throughout the book we will use a superscript asterisk to denote a quantity at equilibrium.

2.1.1 Deterministic Chaos

We can also study a logistic difference equation. Without loss of generality, let us rescale the population abundance in such a way that the maximum carrying capacity is given by $K = 1$. We have

$$x_{t+1} = ax_t(1 - x_t). \tag{2.8}$$

Note that the growth rate in the difference equation, a, is analogous to $1 + r$ in the differential equation (2.6). In contrast to the differential equation, the logistic difference equation (2.8) has many surprises. The behavior of this equation is so rich that many papers and even books have been written about it, and it has generously awarded glorious careers to some scientists who have studied it.

The abundance of the population, x, is given by a number between 0 and 1. The growth rate, a, can vary between 0 and 4. If $a < 0$ or $a > 4$, then negative x values will be generated, which are not biologically meaningful.

The point $x = 0$ is always an equilibrium. If $a < 1$, then the only stable equilibrium of the system is given by $x^* = 0$. This means the population will die out. If $1 < a < 3$, then the only stable equilibrium is given by $x^* = (a - 1)/a$. All trajectories starting from any initial condition x_0 (greater than 0 and less than 1) will converge to this value. The point x^* is a global attractor for the open interval (0, 1).

If $a > 3$, then the point x^* becomes unstable. For a values slightly above 3, we find a stable oscillation of period two. As a increases, the period two oscillator is replaced by period four, then by eight, and so on. For $a = 3.57$ there are infinitely many even periods. For $a = 3.6786$ the first odd periods appear. For $3.82 < a \leq 4$ all periods occur.

The logistic map with $a = 4$ is a simple and most illuminating example for studying deterministic chaos. For any value x_t it is straightforward to compute the population size in the subsequent generation, x_{t+1}. Yet the dynamics are unpredictable in the following sense. Suppose the value of x_t is only known subject to a small uncertainty. It may not be clear whether $x_t = 0.3156$ or 0.3157. After ten generations, however, the trajectories starting from these two initial values will have diverged completely. Hence prediction is impossible. Anything can happen.

We conclude that simple rules can generate complicated behavior. Much of the apparent complexity and unpredictability of biological time series, such as the population size of birds in a particular habitat, the number of measles cases in New York City, or the price fluctuations of stocks and bonds, could in principle be the consequence of deterministic laws.

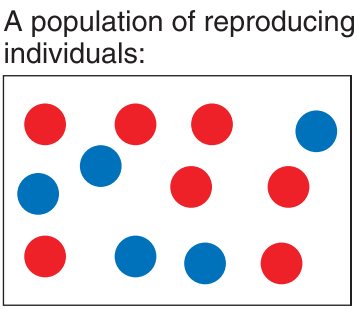

Figure 2.1 Evolution requires a population of reproducing individuals. Strictly speaking, neither genes, nor cells, nor organisms, nor ideas evolve. Only populations can evolve.

2.2 SELECTION

Selection operates whenever different types of individuals reproduce at different rates. At the very least we need two types (Figure 2.1). Let us call them A and B. Type A individuals reproduce at rate a. Type B individuals reproduce at rate b. The rate of reproduction is interpreted as fitness. Therefore the fitness of A is a, the fitness of B is b. Denote by $x(t)$ the number of A individuals at time t. Denote by $y(t)$ the number of B individuals at time t. At time $t = 0$, the numbers of A and B are respectively given by x_0 and y_0. The A and B subpopulations grow according to the differential equations

$$\begin{aligned} \dot{x} &= ax \\ \dot{y} &= by \end{aligned} \tag{2.9}$$

Equation (2.9) is a system of two ordinary, linear differential equations. The analytical solution is given by

$$\begin{aligned} x(t) &= x_0 e^{at} \\ y(t) &= y_0 e^{bt} \end{aligned} \tag{2.10}$$

Hence the A and B subpopulations grow exponentially at rates a and b, respectively. The doubling time for A is $\log 2/a$. The doubling time for B is $\log 2/b$. If a is greater than b, then A reproduces faster than B: after some time, there will be more A than B individuals.

Denote by $\rho(t) = x(t)/y(t)$ the ratio of A over B at time t. We have

$$\dot{\rho} = \frac{\dot{x}y - x\dot{y}}{y^2} = (a - b)\rho. \tag{2.11}$$

The solution of this differential equation, for the initial condition $\rho_0 = x_0/y_0$, is given by

$$\rho(t) = \rho_0 e^{(a-b)t}. \tag{2.12}$$

Hence if $a > b$ then ρ tends to infinity. In this case A will outcompete B, which means selection favors A over B. If, on the other hand, $a < b$, then ρ tends to zero. In this case B will outcompete A, which means that selection favors B over A.

Let us now consider a situation in which the total population size is held constant. This situation can arise, for example, when an ecosystem has a constant maximum carrying capacity. Let $x(t)$ denote the relative abundance of A at time t. Instead of "relative abundance" we can also say "frequency." Let $y(t)$ denote the frequency of B. Since there are only A and B individuals in the population, we have $x + y = 1$. As before, A and B individuals reproduce, respectively, at rates a and b.

We have the system of equations

$$\begin{aligned} \dot{x} &= x(a - \phi) \\ \dot{y} &= y(b - \phi) \end{aligned} \tag{2.13}$$

The term ϕ ensures that $x + y = 1$. This is only possible if $\phi = ax + by$. Observe that ϕ is the average fitness of the population.

The system (2.13) describes only a single differential equation, because y can be replaced by $1 - x$. We obtain

$$\dot{x} = x(1 - x)(a - b). \tag{2.14}$$

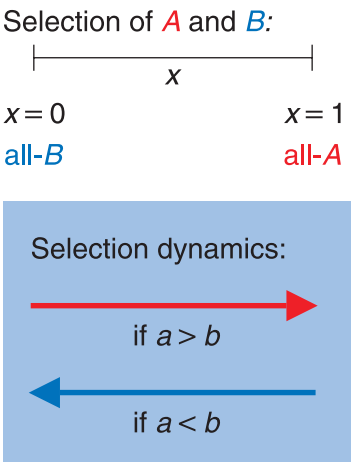

Figure 2.2 Selection arises if two types, A and B, have different rates of reproduction, a and b. If A reproduces faster than B, which means $a > b$, then A will become more abundant than B. Eventually A will take over the entire population; B will become extinct. Denote by x the relative abundance (= frequency) of type A. The quantity x is a number between 0 and 1. Therefore selection dynamics are defined on the closed interval $[0, 1]$.

This differential equation has two equilibria, one for $x = 0$ and the other for $x = 1$. At these two points, we have $\dot{x} = 0$. This observation makes sense: if $x = 1$ then the system consists only of A individuals and nothing more can happen; if $x = 0$, then the system consists only of B individuals and again nothing more can happen.

We can, however, make an additional observation. If $a > b$, then $\dot{x} > 0$ for all values of x that are strictly greater than 0 and strictly smaller than 1. This means that for any mixed system (consisting of some A and some B individuals) the fraction of A will increase if the fitness of A is greater than the fitness of B. In this case, the fraction of B will converge to 0, while the fraction of A converges to 1. We have encountered the concept of "survival of the fitter" (Figure 2.2).

2.2.1 Survival of the Fittest

The model can be extended to describe selection among n different types. Let us label them $i = 1, \ldots, n$. Denote by $x_i(t)$ the frequency of type i. The structure of the population is given by the vector $\vec{x} = (x_1, x_2, \ldots, x_n)$.

Denote by f_i the fitness of type i. As before, fitness is a non-negative real number and describes the rate of reproduction. The average fitness of the

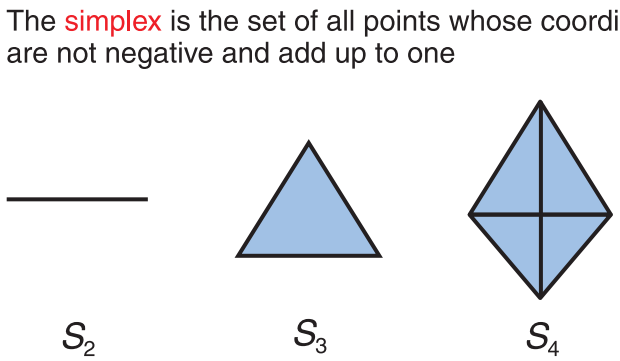

Figure 2.3 If the total population size is constant, then selection dynamics can be formulated in terms of relative abundance (= frequency). Suppose there are n different types, $i = 1, \ldots, n$. Type i has frequency x_i. The sum over all x_i is one. The set of all points, $(x_1, \ldots, x_n)$ with the property $\sum_{i=1}^{n} x_i = 1$, is called the simplex S_n. Selection dynamics occur on the simplex S_n. The figure shows S_2, S_3, and S_4. The simplex S_n is an $n-1$ dimensional structure embedded in an n-dimensional Euclidian space. The simplex S_n has n faces that each consist of the simplex S_{n-1}.

population is given by

$$\phi = \sum_{i=1}^{n} x_i f_i. \tag{2.15}$$

Selection dynamics can be written as

$$\dot{x}_i = x_i(f_i - \phi) \qquad i = 1, \ldots, n \tag{2.16}$$

The frequency of type i increases, if its fitness exceeds the average fitness of the population. Otherwise it will decline. The total population size remains constant: $\sum_{i=1}^{n} x_i = 1$ and $\sum_{i=1}^{n} \dot{x}_i = 0$.

The set of points with the property $\sum_{i=1}^{n} x_i = 1$ is called the simplex S_n (Figure 2.3). Each point in the simplex refers to a particular structure of the population. The interior of the simplex is the set of points $\vec{x}$ with the property that $x_i > 0$ for all $i = 1, \ldots, n$. The face of the simplex is the set of points $\vec{x}$ with the property that $x_i = 0$ for at least one i. The vertices of the simplex

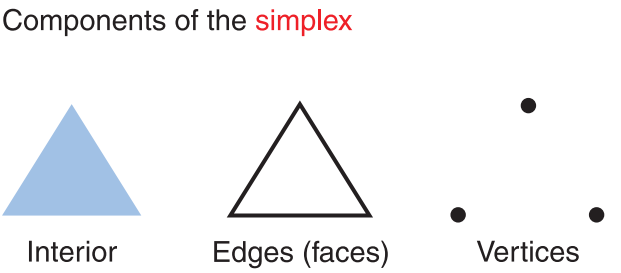

Figure 2.4 The *interior* of a simplex is the set of all points where all coordinates are strictly positive; this means no type has become extinct. The *faces* are the sets of points where at least one coordinate is zero; this means at least one type has become extinct. The *vertices* describe pure populations, where all but one type have become extinct.

are the corner points where exactly one type is present, $x_i = 1$, while all other types are extinct, $x_j = 0$ for all $j \neq i$ (Figures 2.4 and 2.5).

The simplex S_2 is given by the closed interval [0, 1]. The notation [0, 1] refers to all numbers which are greater than or equal to 0 and less than or equal to 1. In contrast, (0, 1) is the open interval; it contains all numbers that are strictly greater than 0 and strictly less than 1. The open interval (0, 1) is the interior of the closed interval [0, 1] and, therefore, is also the interior of the simplex S_2.

Equation (2.16) contains a single globally stable equilibrium. Starting from any initial condition in the interior of the simplex, the population will converge to a corner point where all but one type have become extinct. The winner, k, enjoys a well-deserved victory because it has the property of having the largest fitness, f_k. Thus $f_k > f_i$ for all $i \neq k$. The system shows competitive exclusion: the fittest type will outcompete all others. This is the concept of "survival of the fittest."

2.2.2 Survival of the First, Survival of All

Let us return to the selection of two types, A and B, but without making the assumption that their growth rates are linear functions of their frequencies. Instead consider the equation

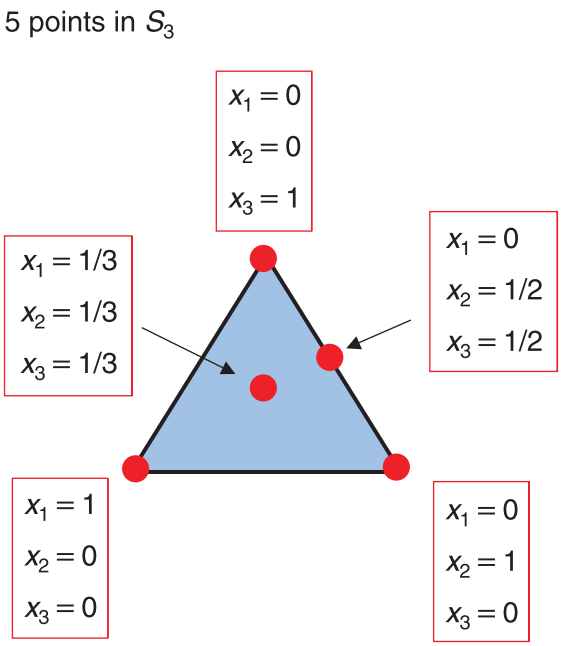

Figure 2.5 Five points on the simplex S_3. In the center, (1/3, 1/3, 1/3), all three types have the same frequency. There are three faces. The center of one particular face is given by (0, 1/2, 1/2); one type has become extinct. The corner points (vertices) indicate populations that consist of only one type. S_3 has three corners: (1, 0, 0), (0, 1, 0), and (0, 0, 1).

$$\begin{aligned} \dot{x} &= ax^c - \phi x \\ \dot{y} &= by^c - \phi y \end{aligned} \tag{2.17}$$

As before, a and b denote the fitness values of A and B, respectively. If $c = 1$, we are back to equation (2.13). If $c < 1$, then growth is subexponential. In the absence of the density limitation, ϕ, the growth curve of the two types would be slower than exponential.

In contrast, if $c > 1$, then growth is superexponential. In the absence of the density limitation, ϕ, the growth curve of the two types would be faster than exponential (hyperbolic). To maintain a constant population size, $x + y = 1$, we set $\phi = ax^c + by^c$. Equation (2.17) reduces to

$$\dot{x} = x(1 - x)f(x) \tag{2.18}$$

where

$$f(x) = ax^{c-1} - b(1 - x)^{c-1}. \tag{2.19}$$

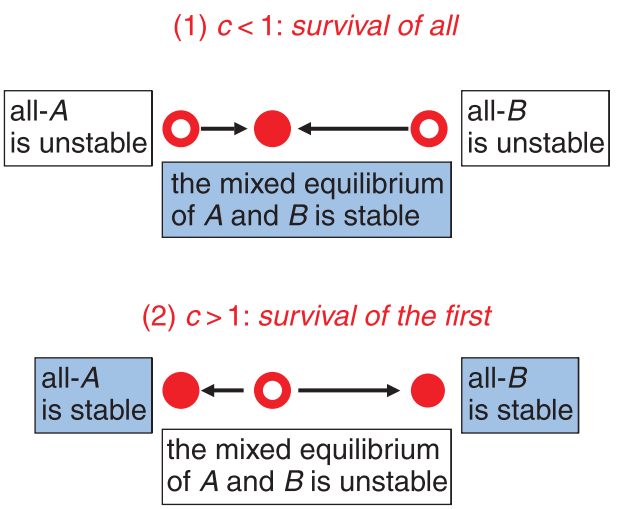

Figure 2.6 Survival of all: for subexponential growth ($c < 1$), there is a stable mixed equilibrium between A and B, even if one type has a faster growth rate than the other. Survival of the first: for superexponential growth ($c > 1$), there is an unstable mixed equilibrium between A and B, while the pure populations are stable. For example, if the whole population consists of B, then A cannot invade even if it has a higher growth rate.

This equation always has fixed points for $x = 0$ and $x = 1$. For $c \neq 1$ there exists exactly one other fixed point between 0 and 1. It is given by

$$x^* = \frac{1}{1 + \sqrt[c-1]{a/b}}. \tag{2.20}$$

If $c < 1$, then the boundary fixed points, $x = 0$ and $x = 1$, are always unstable; the interior fixed point, x^*, is globally stable. Hence there is survival of both A and B. Surprisingly, even if A is fitter than B, $a > b$, then a small amount of B can invade an A population.

If $c > 1$, then the boundary fixed points, $x = 0$ and $x = 1$, are always stable; the interior fixed point, x^*, is unstable. If $x > x^*$, then A will outcompete B. If $x < x^*$, then B will outcompete A. Again this observation is remarkable. Even if A is fitter than B in the sense that $a > b$, a B population cannot be invaded by an A mutant.

We conclude that superexponential growth favors whoever was there first (survival of the first) whereas subexponential growth leads to the survival of all (Figure 2.6).

What is the intuition behind this observation? An extreme form of subexponential growth is "immigration," $c = 0$. The growth rate does not depend on x or y at all. We have

$$\begin{aligned} \dot{x} &= a - \phi x \\ \dot{y} &= b - \phi y \end{aligned} \tag{2.21}$$

with $\phi = a + b$. This equation can be interpreted as the immigration of A and B into the population from some other place. It is clear that these immigration dynamics lead to coexistence. A value of c between 0 and 1 is a mixture between immigration and linear growth and retains the property of coexistence.

If $c > 1$, on the other hand, then A cannot invade B even if $a > b$. ("Invasion" means that an infinitesimally small fraction of A individuals can increase in abundance in a population where almost all individuals are of type B.) The intuitive explanation is as follows: we can think of the case $c = 2$ as implying that two individuals of the same type have to meet in order to reproduce. If there is only an infinitesimally small fraction of A individuals, then two A individuals will never meet and hence A will not reproduce. If $c = 3$ then three individuals of the same type have to meet in order to reproduce. Again arbitrarily small fractions of a type can never increase. The same intuition holds for all values $c > 1$.

The case $c = 2$ can also be interpreted as an evolutionary game between two strategies, A and B, that are strict Nash equilibria. Neither strategy can invade the other. We will encounter these concepts in Chapter 4.

2.3 MUTATION

Life takes advantage of mistakes. Replication of DNA or RNA can lead to slightly modified sequences that represent novel variants. Errors during reproduction are called mutations. In this section, we study the simplest possible differential equations that describe mutation (Figure 2.7).

Let us again consider just two types, A and B. Denote by u_1 the mutation rate from A to B: u_1 is the probability that the reproduction of A leads to B.

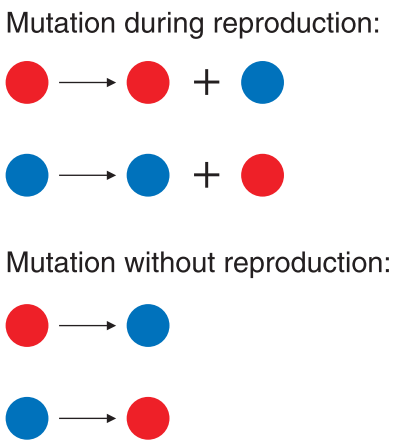

Figure 2.7 Mutation can occur during reproduction: type A produces an offspring that is type B. Mutation can also occur in the absence of reproduction: type A changes into type B. Many genetic mutations occur when the genomic material of a cell is being copied. But mutagens can also change the genetic material of a cell when it is not dividing.

Conversely, denote by u_2 the mutation rate from B to A. As before, let x and y denote the frequencies of A and B, respectively. We have

$$\begin{aligned} \dot{x} &= x(1-u_1) + yu_2 - \phi x \\ \dot{y} &= xu_1 + y(1-u_2) - \phi y \end{aligned} \tag{2.22}$$

Since A and B have the same fitness ($a = b = 1$), the average fitness of the population is constant and given by $\phi = 1$. Taking into account $x + y = 1$, system (2.22) reduces to the differential equation

$$\dot{x} = u_2 - x(u_1 + u_2). \tag{2.23}$$

The frequency of A converges to the stable equilibrium

$$x^* = \frac{u_2}{u_1 + u_2}. \tag{2.24}$$

Hence mutation leads to coexistence between A and B. The relative proportion of A and B at equilibrium depends on the mutation rates. At equilibrium, the ratio of A to B is given by $x^*/y^* = u_2/u_1$. If the mutation rates are the same, $u_1 = u_2$, and then $x^* = y^*$.

Sometimes the mutation rate in one direction is much larger than in the other direction. In these cases, it often makes sense to ignore mutation in the other direction altogether. Let $u_2 = 0$. We have

$$\dot{x} = -xu_1. \tag{2.25}$$

Therefore the frequency of A declines over time as

$$x(t) = x_0 e^{-u_1 t}. \tag{2.26}$$

The frequency of B increases as

$$y(t) = 1 - (1 - y_0)e^{-u_1 t}. \tag{2.27}$$

If mutation occurs only from A to B but not the other way around, then A will die out and B will take over the whole population. We see that mutation can affect survival. Different mutation rates can introduce selection even in the absence of different reproductive rates.

2.3.1 Mutation Matrix

We can extend mutation dynamics to n different types. Let us introduce the mutation matrix, $Q = [q_{ij}]$. The probability that type i mutates to type j is given by q_{ij}. Since each type i has to produce itself or some other type, we have $\sum_{j=1}^{n} q_{ij} = 1$. Thus Q is a stochastic $n \times n$ matrix. A stochastic matrix is defined by the properties that (i) all entries are numbers from the interval $[0, 1]$ (so-called probabilities), (ii) there are as many rows as columns, and (iii) the sum of each row is 1. Stochastic matrices always have 1 as an eigenvalue, and no eigenvalue has an absolute value greater than 1.

Mutation dynamics can be written as

$$\dot{x}_i = \sum_{j=1}^{n} x_j q_{ji} - \phi x_i \qquad i = 1, \ldots, n \tag{2.28}$$

In vector notation we can write

$$\dot{\vec{x}} = \vec{x}Q - \phi\vec{x}. \tag{2.29}$$

Again the average fitness is just $\phi = 1$. The equilibrium is given by the left-hand eigenvector associated with eigenvalue 1:

$$\vec{x}^* Q = \vec{x}^*. \tag{2.30}$$

The point $\vec{x}^*$ denotes the unique globally stable equilibrium of the mutation dynamics.

2.4 MATING

One of the problems that Charles Darwin could not solve was the following: under random mating and blending inheritance, the variability in a population should rapidly decline. Yet it was clear that variability was needed for natural selection. If variability disappears, then natural selection has nothing upon which to act. Suppose there is a distribution of body size in a population. If children inherit the average body size of their parents, then after some time everybody is the same size. Under these circumstances, how can natural selection affect changes in body size?

The first part of the solution is that inheritance (on the level of genes) is not blending but particulate, as had been discovered by Gregor Mendel and published in 1866. That is, individuals have discrete genotypes that get reshuffled, not blended, during mating. Mendel's work was unknown to Darwin. The second step was a simple mathematical analysis, which was performed by the British mathematician G. H. Hardy, who was proud never to have done anything useful (= applied) in his life, only to have his name forever associated with a highly useful and very applied concept in population genetics. Moreover, Hardy's brief calculation was generalized by the German physician Wilhelm Weinberg.

Consider an infinitely large population of a diploid organism with two sexes and random mating (a diploid organism has two copies of its genome; humans and many other animals are diploid). Let us look at one particular gene locus and assume there are two alleles, A_1 and A_2. The alleles are variants of the same gene and might differ in one or a few point mutations. (Point mutation means that only one single base of the DNA sequence is changed.)

There are 3 different genotypes: A_1A_1, A_1A_2, A_2A_2. Let us denote their frequencies in the population by x, y, and z, respectively. Denote by p and q the frequencies of alleles A_1 and A_2. We have $x + y + z = 1$ and $p + q = 1$. Moreover,

$$\begin{aligned} p &= x + \frac{1}{2}y \\ q &= z + \frac{1}{2}y \end{aligned} \tag{2.31}$$

Let us now assume random mating. In the next generation, the genotype frequencies are given by

$$\begin{aligned} x' &= p^2 \\ y' &= 2pq \\ z' &= q^2 \end{aligned} \tag{2.32}$$

For the allele frequencies in the next generation we have again

$$\begin{aligned} p' &= x' + \frac{1}{2}y' \\ q' &= z' + \frac{1}{2}y' \end{aligned} \tag{2.33}$$

Combining (2.32) and (2.33), we observe that

$$p' = p \qquad q' = q \tag{2.34}$$

Therefore the allele frequencies remain unchanged from one generation to the next. Moreover, combining (2.32) and (2.34), we observe

$$\begin{aligned} x' &= p'^2 \\ y' &= 2p'q' \\ z' &= q'^2 \end{aligned} \tag{2.35}$$

From the first generation on, the genotype frequencies can be directly derived from the allele frequencies. Note that equation (2.35) need not hold for the initial genotype and allele frequencies. The Hardy-Weinberg law (expressed by equations 2.34 and 2.35) can be generalized to n alleles.

In summary, the Hardy-Weinberg law states that particulate inheritence preserves variation within a population under random mating.

SUMMARY

- Evolution requires populations of reproducing individuals.
- Asexual reproduction leads to exponential population growth (which will eventually be checked by resource limitation).
- Simple models of population growth in discrete time can give rise to very complicated dynamics.
- Selection arises when different types of individuals reproduce at different rates.
- Normally, the faster-reproducing (fitter) individual outcompetes the slower reproducing (less fit) individual.
- If there are many different types, then selection dynamics can lead to "survival of the fittest." All others become extinct.
- Sublinear growth rates lead to coexistence, "survival of all."
- Superlinear growth rates prevent invasion of a new type and thereby lead to "survival of the first."
- Mutation arises when reproduction is not perfectly accurate.
- Mutation promotes coexistence of different types.
- Asymmetric mutation can lead to selection even if all individuals have the same reproduction rate.
- The Hardy-Weinberg law states that random mating preserves genetic variation within a population.

3 FITNESS LANDSCAPES AND SEQUENCE SPACES

GENOMES ARE SEQUENCES of the four-letter alphabet A, T, C, G, denoting the nucleotides adenine, thymine, cytosine, and guanine. All living cells use double-stranded DNA to carry their genomic information. Many viruses also use DNA, but some viruses encode their genome in form of RNA. The genome length of organisms varies greatly, ranging from about 10^4 nucleotides for small viruses, to 10^6 for bacteria to 3×10^9 for humans. Curiously, newts and lungfish "need" an even larger genome than do humans (19×10^9 and 140×10^9, respectively). The evolutionary dynamics of genome size and genome organization is a fascinating topic.

If a cell wants to produce a particular protein, then the DNA of the corresponding gene is "transcribed" into messenger RNA (mRNA), which is in turn "translated" into protein. The transcription is done by particular enzymes called DNA-dependent RNA polymerases. The translation is performed by a complicated arrangement of RNA and proteins called ribosomes. The words "transcription" and "translation" were invented by the mathematician John von Neumann when he calculated how to build a self-reproducing machine. He came up with an architecture equivalent to the organization of cells some decades before molecular biology had been invented.

RNA also uses a four-letter alphabet, A, U, C, G. Thymine is replaced by uracil. Furthermore, the sugar backbone of RNA has an additional -OH (hydroxy) group, which makes the molecule less stable and more dynamic. DNA is a stable carrier of information. RNA also carries information, but in addition some RNAs have enzymatic activity.

Proteins consist of 20 amino acids. Each amino acid is encoded by a sequence of three letters of the RNA alphabet. This genetic code is essentially the same for all living cells, ranging from bacteria to humans to newts. Hence the genetic code is believed to have originated only once: in the first cell that is ancestor to all existing cells. A 4-letter alphabet generates 64 possible sequences of length 3. Since there are only 20 amino acids, the genetic code is redundant: some amino acids are encoded by more than one sequence. Some sequences are used to signal the end of the transcription process. We see that molecular biology adds a precise information-theoretic perspective to evolutionary dynamics.

3.1 SEQUENCE SPACE

In the green hills of Sussex lived an imaginative theoretical biologist, John Maynard Smith, who once pictured all proteins (of a certain length) arranged in such a way that nearest neighbors differed by a single amino acid. This was the origin of what we call "sequence space" as a concept in the human mind.

Let us consider all proteins of the modest length 100. Each position of the protein sequence is filled by one of 20 amino acids. Hence, this space has 100 dimensions and in total 20^{100} points. This number corresponds to 10^{130} proteins. In contrast, there are only some 10^{80} estimated protons in our universe. Nor is there any reason for us to consider only proteins of length 100; some proteins are much longer. We conclude there are many more possible proteins than available protons and hence evolution so far has and, for the remaining 10^{30} years that constitute the lifetime of our protons, will only explore a vanishingly small subset of all possible proteins.

What is true of proteins is also true of genes and genomes. We can imagine all nucleotide sequences of a certain length arranged in a way that nearest neighbors differ in one position. For sequence length L this generates a lattice

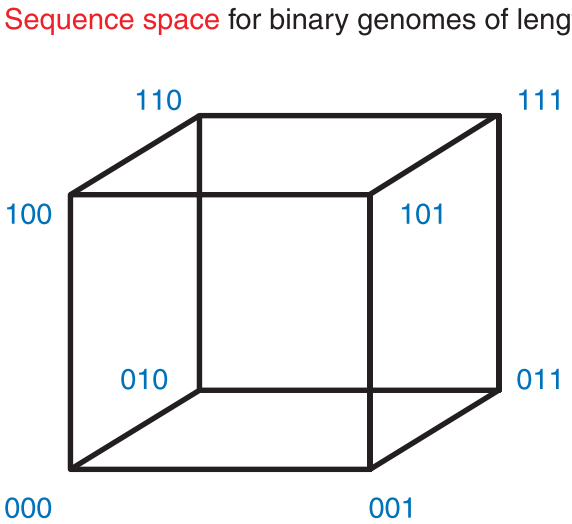

Figure 3.1 Genomes live in sequence space. The number of dimensions is given by the length of the genome. Small viruses live in 10,000 dimensions. Humans live in about 3 billion dimensions.

in an L-dimensional space. In each dimension there are 4 discrete possibilities. Hence there are 4^L possible sequences.

For writing computer programs, it is often convenient to use binary sequences, the fundamental strings of silicon thoughts. Moreover, everything from Shakespeare to *E. coli* can be encoded in binary sequences. For length L there are 2^L possibilities. In Figure 3.1, the binary sequence space for $L = 3$ is shown. The distance between 000 and 010 is one. The distance between 000 and 011 is 2 (and not $\sqrt{2}$). Hence sequence space is characterized not by a Euclidean metric but by a so-called Hamming metric or Manhattan metric. In Manhattan, if you are on 5th Avenue and 51st Street it takes 2 blocks to go to 6th Avenue and 52nd Street, not $\sqrt{2}$ blocks. This metric was introduced by Richard Hamming in information theory.

Let us compare the binary sequence space of length $L = 300$ with a three-dimensional cubic lattice containing the same number of points. There are

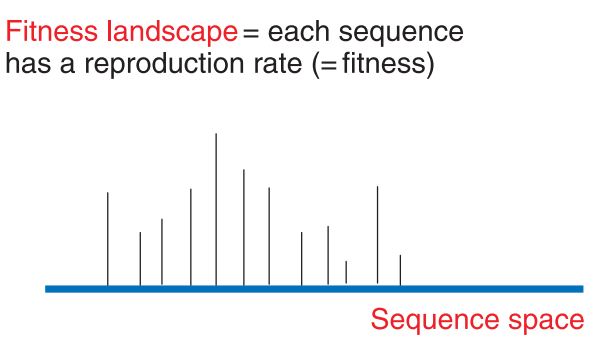

Figure 3.2 The fitness landscape is a high-dimensional mountain range. Each genome (= each point in sequence space) gets assigned a fitness value.

$2^{300} \approx 10^{90}$ points. Imagine nearest neighbors are placed at a distance of 1 meter. The diagonal of the three dimensional cubic lattice has a length of about 10^{30} meters, which corresponds to about 10^{14} light years. In contrast, the longest distance in the L-dimensional hypercube is only 300 meters. Thus sequence space is characterized by short distances, but many dimensions. It is not far to move from one sequence to another, but there are many possible steps that lead in wrong directions. Evolution is a trajectory through sequence space. This trajectory needs an efficient guide.

3.2 FITNESS LANDSCAPES

The American population geneticist Sewall Wright invented the concept of a "fitness landscape" in the 1930s, but Manfred Eigen and Peter Schuster, collaborating in the 1970s, combined fitness landscape with sequence space. Consider a function that assigns to each genomic sequence a fitness value. Hence we build a mountain range on the foundation of an L-dimensional sequence space (Figure 3.2). This mountain range has $L + 1$ dimensions. The evolutionary process of mutation and selection explores this hyper-alpine mountain range.

The genomic sequence represents the genotype of an organism. The phenotype of an organism is given by its shape, behavior, performance and any kind of ecological interaction. The phenotype determines the fitness (reproductive rate) of the organism. There is a mapping from genotype to phenotype.

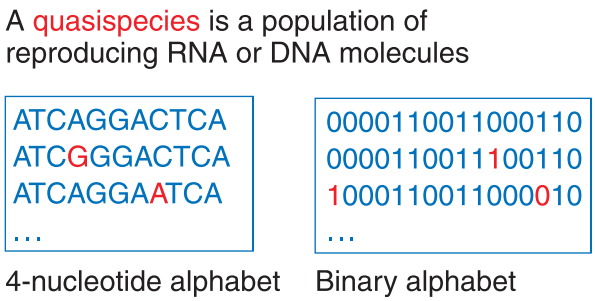

Figure 3.3 The ensemble of genomes of a natural population form a quasispecies: the genomes of different individuals are similar but not identical. Biology has chosen a four-letter alphabet consisting of the nucleotides A, T, C, and G for its genes. Most in silico evolution uses a binary alphabet for convenience. Sequence differences (mutations) are shown in red.

There is another mapping from phenotype to fitness. The fitness landscape is a convolution of these two mappings. It is a direct mapping from genotype to fitness.

The fitness landscape of certain problems can be determined experimentally. For example, HIV can generate point mutations that confer drug resistance. The relative growth rate of such mutants can be determined by in-vitro assays. In general, however, to understand the relationship between genotype, phenotype, and fitness is an extremely complicated problem. Much of biology, including developmental biology, molecular biology, post-genomics, and proteomics, is devoted to this very task.

3.3 THE QUASISPECIES EQUATION

A quasispecies is an ensemble of similar genomic sequences generated by a mutation-selection process (Figure 3.3). The term was introduced by the chemists Manfred Eigen and Peter Schuster. In chemistry the word "species" refers to an ensemble of identical molecules, for example, the species of all water molecules. But the species of all RNA molecules does not contain identical sequences, and therefore the term "quasispecies" was coined. Biologists

are sometimes confused by this expression, because they relate it to the concept of a biological species.

We stay with binary sequences for convenience. We note that any genomic or other information can be encoded by binary sequences. Consider all binary sequences of length L. Enumerate all those sequences by $i = 0, 1, 2, \ldots, n$ where $n = 2^L - 1$. A natural enumeration is obtained if the sequence represents the binary description of the corresponding integer. For example, let $L = 4$. The sequence 0000 corresponds to $i = 0$, the sequence 0001 to $i = 1$, the sequence 0010 to $i = 2, \ldots$, the sequence 1111 to $i = 15$.

Imagine an infinitely large population of organisms, each carrying a genome of length L. Denote by x_i the relative abundance (= frequency) of those organisms that contain genome i. We have $\sum_{i=0}^{n} x_i = 1$. The genomic structure of the population is given by the vector $\vec{x} = (x_0, x_1, \ldots, x_n)$.

Denote by f_i the fitness of genome i. It is a non-negative real number. Thus genomes of type i are being reproduced at rate f_i. The fitness landscape is given by the vector $\vec{f} = (f_0, f_1, \ldots, f_n)$. The average fitness of the population, $\phi = \sum_{i=0}^{n} x_i f_i$, is the inner product of the vectors $\vec{x}$ and $\vec{f}$. We have $\phi = \vec{x}\vec{f}$.

During replication of a genome, mistakes can happen. The probability that replication of genome i results in genome j is given by q_{ij}. Here we again meet the mutation matrix $Q = [q_{ij}]$ of section 2.3. We remember that Q is a stochastic matrix: it has as many rows as columns; each entry is a probability, which means a number between 0 and 1; each row sums to one, $\sum_{j=0}^{n} q_{ij} = 1$.

The quasispecies equation (Figure 3.4) is given by

$$\dot{x}_i = \sum_{j=0}^{n} x_j f_j q_{ji} - \phi x_i \qquad i = 0, \ldots, n \tag{3.1}$$

Sequence i is obtained by replicating any sequence j at rate f_j times the probability that replication of sequence j generates sequence i. Each sequence is removed at rate ϕ to ensure that the total population size remains constant, $\sum_{i=0}^{n} x_i = 1$. Thus quasispecies dynamics are defined on the simplex, S_n.

In the limiting case of completely error-free replication, Q becomes the identity matrix: all diagonal entries are one, all off-diagonal entries are zero.

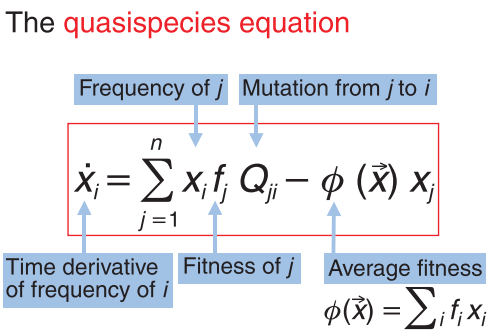

Figure 3.4 The quasispecies equation, formulated by Manfred Eigen and Peter Schuster, is one of the most important equations in theoretical biology. It describes the mutation and selection of an infinitely large population on a constant fitness landscape.

Consider an initial condition in the interior of the simplex, defined by $x_i > 0$ for all i. The quasispecies will converge to a homogeneous population that consists only of the fittest sequence. If $f_0 > f_i$ for all $i \neq 0$, then the stable equilibrium is given by $x_0 = 1$ and $x_i = 0$ for $i \neq 0$. If there are no errors, then the quasispecies equation (3.1) reduces to the selection equation (2.16) of section 2.2.1.

Let us now assume that errors occur. This means that (at least some) off-diagonal entries of Q are not zero. In many realistic contexts, the matrix Q is irreducible, which means it is possible to find a sequence of mutations from any one genome i to any other genome j. Furthermore, let $f_i > 0$ for at least some i. In this case, the quasispecies equation admits a single, globally stable equilibrium, $\vec{x}^*$, in the simplex S_n.

The equilibrium quasispecies, $\vec{x}^*$, does not necessarily maximize the average fitness ϕ. Consider again a fitness landscape with the property $f_0 > f_i$ for all $i \neq 0$. Then the population consisting only of sequence 0 will have a higher fitness than the equilibrium population $\vec{x}^*$. Thus, mutations reduce the average fitness at equilibrium.

Observe that (3.1) is a nonlinear differential equation. The term $-\phi x_i$ is of second order. Linear differential equations can always be solved, but nonlinear differential equations normally cannot be solved. This means for nonlinear differential equations the trajectories cannot always be written as explicit

functions of time. The quadratically nonlinear quasispecies equation (3.1), however, can be solved as follows. First, define

$$\psi(t) = \int_0^t \phi(s)ds. \tag{3.2}$$

Note that

$$\dot{x}_i + \phi x_i = e^{-\psi} \frac{d(x_i e^{\psi})}{dt}. \tag{3.3}$$

Let us define

$$X_i(t) = x_i(t)e^{\psi(t)}. \tag{3.4}$$

Now $X_i(t)$ is given by the linear equation

$$\dot{X}_i = \sum_{j=0}^{n} X_j f_j q_{ji} \qquad i = 0, \ldots, n \tag{3.5}$$

This system of linear differential equations describes exponential growth of all the members of the quasispecies. The linear system (3.5) can be solved using standard techniques. Notice also that

$$X = \sum_{i=0}^{n} X_i = \left(\sum_{i=0}^{n} x_i \right) e^{\psi} = e^{\psi}. \tag{3.6}$$

This means, from equation (3.4), that we can write $x_i = X_i / X$, which in turn means that X_i can be interpreted as the absolute abundance of individuals with genome i. Also note that X, the total population size, grows as

$$\dot{X} = \dot{\psi} e^{\psi} = \phi X. \tag{3.7}$$

Therefore the total population size grows exponentially at a rate that is given by the average fitness, ϕ, of the population.

Let us combine the fitness landscape, $\vec{f}$, and the mutation matrix, Q, to obtain the mutation-selection matrix,

$$W = [w_{ji}] = [f_j q_{ji}]. \tag{3.8}$$

Quasispecies dynamics are determined by the properties of the matrix, W. In vector notation the quasispecies equation can be written as

$$\dot{\vec{x}} = \vec{x}W - \phi\vec{x}. \tag{3.9}$$

Hence the equilibrium of quasispecies dynamics is given by

$$\vec{x}W = \phi\vec{x}. \tag{3.10}$$

This is a standard eigenvalue problem. The average fitness, ϕ, is the largest eigenvalue of the matrix W. The left-hand eigenvector associated with this eigenvalue, with the proper normalization $\sum_{i=1}^{n} x_i = 1$, provides the equilibrium structure of the quasispecies. Generically, there is a unique and globally stable equilibrium.

3.4 A MUTATION MATRIX FOR POINT MUTATIONS

During the replication of a DNA or RNA genome, many types of mutational events can occur. "Point mutations" describe the change of one base for another. "Insertions" denote the addition of a string of bases to the existing sequence. "Deletions" characterize the reverse process, the loss of a string of bases. "Recombination" means that genetic material can be exchanged between two sequences. Here we will only deal with point mutations of binary sequences.

Let us consider the set of all sequences of a given length, L. The Hamming distance, h_{ij}, counts the number of positions that differ between sequences i and j. For example, the Hamming distance between the sequences 1010 and 1100 is two. Denote by u the probability that a mutation occurs in a specific position. Thus $1 - u$ is the probability that the mutation is copied correctly. We can write the probability that replication of sequence i results in sequence j as

$$q_{ij} = u^{h_{ij}}(1-u)^{L-h_{ij}}. \tag{3.11}$$

Hence a mutation has to occur in as many positions as differ between the sequences i and j, which is precisely the Hamming distance, h_{ij}. No mutation must occur in the remaining $L - h_{ij}$ positions.

Equation (3.11) is an elegant description of a mutation matrix that allows point mutations among binary sequences of constant length. It is assumed that the point mutation rate, u, is the same for all positions. It is further assumed that a mutation in one position is independent of a mutation in another position. Hence one error does not increase the probability of another error. There are no insertions and no deletions. All of these restrictions can be relaxed in principle, but doing so will lead to considerable complexity.

Let us use mutation matrix (3.11) to describe the human immunodeficiency virus as an example. The point mutation rate of HIV is approximately $u = 3 \times 10^{-5}$. The genome length of HIV is $L = 10^4$. Therefore the probability that the whole HIV genome is replicated without mutation is given by $(1 - u)^L \approx 0.74$. The probability that replication of the HIV genome results in a sequence that differs in one arbitrary position is given by $Lu(1 - u)^{L-1} = 0.22$. The probability that a particular one-error mutant, for example one that confers drug resistance or immune escape, is being produced is given by $u(1 - u)^{L-1} = 2.2 \times 10^{-5}$. If 10^9 newly infected cells are being produced each day, then any particular one-error mutant will arise 22,000 times each day. This number signifies the enormous potential of HIV (or other viruses or microbes) to escape from selection pressures that are meant to control them. We will revisit this topic in Chapter 10.

3.5 ADAPTATION IS LOCALIZATION IN SEQUENCE SPACE

The quasispecies equation (3.1) describes the movement of a population through sequence space. The quasispecies "feels" gradients in the mountain range of the fitness landscape. It attempts to climb uphill and reach local or global peaks (Figure 3.5). What are the conditions that this evolutionary walk will be successful? One such condition is the error threshold.

If the mutation rate u is too high, then the ability of the quasispecies to climb uphill and to remain on top of a mountain peak is impaired. In fact, we can show that for many natural fitness landscapes there is a maximum mutation rate, u_c, that is still compatible with adaptation. If the mutation rate exceeds this value, $u > u_c$, then adaptation is not possible.

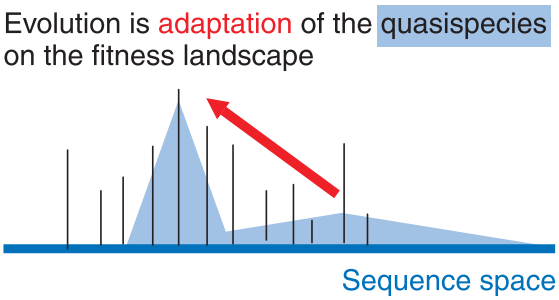

Figure 3.5 Quasispecies love to climb mountains in high-dimensional spaces. The higher they get, the fitter they are. Adaptation means to go up.

Adaptation means that the quasispecies is able to find peaks in the fitness landscape and stay there. Suppose the fitness landscape contains only one peak. If the mutation rate is sufficiently low, then the equilibrium solution of equation (3.1) describes a quasispecies that is centered on this peak. Most sequences resemble the type with maximum fitness or nearby mutants. Sequences that are far away from the peak will have a very low frequency. (In population genetics, frequency means relative abundance.) We say the quasispecies is adapted to this peak. Similarly, we can say that the quasispecies distribution is localized at this peak. Adaptation means localization in sequence space. When the mutation rate of a quasispecies is zero, it contains only sequences with maximum fitness. When the mutation rate is very small, the quasispecies distribution is very narrow. As the mutation rate increases, the quasispecies distribution widens. There is a critical mutation rate, u_c, beyond which the equilibrium quasispecies no longer "feels" the peak. The quasispecies is no longer localized around the peak. Adaptation is lost. Strictly speaking, a well-defined "phase transition" from a localized to a delocalized state only occurs for infinite sequence length, but the phenomenon is striking already for binary sequences of length $L = 10$.

The maximum mutation rate, u_c, that is compatible with adaptation is called the "error threshold." Not all fitness landscapes have error thresholds. Narrow peaks of finite height have error thresholds. If a peak is so broad that most sequences in the sequence space are within the slopes of the peak, then an error threshold need not occur.

Quasispecies have a tendency to climb uphill. Starting from some random initial condition, $\vec{x}(0)$, the quasispecies equation (3.1) will tend to increase the average fitness, ϕ. But it is also easy to construct a counterexample. Suppose a certain sequence has maximum fitness, while all other sequences have lower fitness. If we start with a population that contains only the sequence with maximum fitness, then equation (3.1) will reduce the average fitness ϕ until an equilibrium between mutation and selection, a so-called mutation-selection balance has been reached.

Calculating the error threshold, u_c, for complex fitness landscapes is difficult, but the following simple fitness landscape provides the crucial insight. Consider all binary sequences of length L. The all-zero sequence, $00\ldots0$, has the highest fitness given by $f_0 > 1$. All other sequences have fitness 1. The all-zero sequence is sometimes called the "master sequence" or the wild type, while all other sequences are called "mutants."

The probability that the master sequence produces an exact copy of itself is given by $q = (1-u)^L$. The probability that the master sequence generates any mutant is given by $1-q$. The trick is to neglect the back mutation from the mutants to the master sequence. With this assumption the quasispecies equation (3.1) becomes

$$\begin{aligned} \dot{x}_0 &= x_0(f_0 q - \phi) \\ \dot{x}_1 &= x_0 f_0 (1-q) + x_1 - \phi x_1 \end{aligned} \tag{3.12}$$

Here x_0 is the frequency of the master sequence, while x_1 is the sum of the frequencies of all the mutants. Clearly, $x_0 + x_1 = 1$. The average fitness is given by $\phi = f_0 x_0 + x_1$. System (3.8) collapses to a single equation

$$\dot{x}_0 = x_0[f_0 q - 1 - x_0(f_0 - 1)]. \tag{3.13}$$

If $f_0 q < 1$, then x_0 will converge to zero; the fittest sequence cannot be maintained in the population. If $f_0 q > 1$, then x_0 will converge to

$$x_0^* = \frac{f_0 q - 1}{f_0 - 1}. \tag{3.14}$$

Hence, the error threshold is given by

$$f_0 q > 1. \tag{3.15}$$

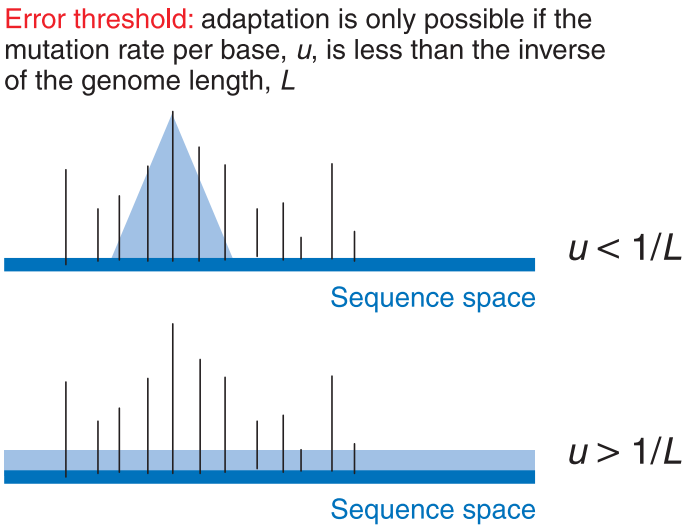

Figure 3.6 Error threshold: a quasispecies can only maintain a peak in a fitness landscape if the mutation rate is less than the inverse of the genome length. This is a very general and beautiful result that must hold for any living organism. The beauty is not spoilt by two qualifying remarks that are necessary: (i) the genome length, L, has to be defined properly to include only those positions that affect fitness and (ii) there are some pathological landscapes where a peak can be maintained beyond the error threshold, for example if the peak is "infinitely" high or so wide that its presence can be felt by the majority of all possible sequences.

This inequality can be rewritten as $\log f_0 > -L \log(1-u)$. For small mutation rates, $u \ll 1$, we have $\log(1-u) \approx -u$. Therefore we obtain the condition

$$u < \frac{\log f_0}{L}. \tag{3.16}$$

If the fitness advantage of the master sequence is not too large and not too small, then $\log f_0$ is approximately 1. Now the error-threshold condition reduces to

$$u < 1/L. \tag{3.17}$$

Hence the maximum mutation rate that is still compatible with adaptation has to be less than the inverse of the genome length (Figure 3.6). In other

Table 3.1 Genome length (in bases), mutation rate per base, and mutation rate per genome for organisms ranging from DNA viruses to humans

Organism	Genome length in bases	Mutation rate per base	Mutation rate per genome
RNA viruses			
Lytic viruses			
Qβ	4.2×10^3	1.5×10^{-3}	6.5
Polio	7.4×10^3	1.1×10^{-4}	0.84
VSV	1.1×10^4	3.2×10^{-4}	3.5
Flu A	1.4×10^4	7.3×10^{-6}	0.99
Retroviruses			
SNV	7.8×10^3	2.0×10^{-5}	0.16
MuLV	8.3×10^3	3.5×10^{-6}	0.029
RSV	9.3×10^3	4.6×10^{-5}	0.43
Bacteriophages			
M13	6.4×10^3	7.2×10^{-7}	0.0046
λ	4.9×10^4	7.7×10^{-8}	0.0038
T2 and T4	1.7×10^5	2.4×10^{-8}	0.0040
E. coli	4.6×10^6	5.4×10^{-10}	0.0025
Yeast *(S. cerevisiae)*	1.2×10^7	2.2×10^{-10}	0.0027
Drosophila	1.7×10^8	3.4×10^{-10}	0.058
Mouse	2.7×10^9	1.8×10^{-10}	0.49
Human *(H. sapiens)*	3.5×10^9	5.0×10^{-11}	0.16

Sources: Drake (1991, 1993) and Drake et al. (1998).
Note: Most organisms have a mutation rate per genome which is less than one, as predicted by the error threshold theory. Why Qβ and VSV have such a high mutation rate is at present unexplained.

words, the genomic mutation rate, uL, has to be less than one. In fact, this condition holds for most living organisms for which mutation rates have been measured (Table 3.1). For eukaryotes, the genome length L in this context should actually be defined as the total number of bases in the coding and regulatory regions of the DNA.

3.6 SELECTION OF THE QUASISPECIES

The following remarkable observation was first made by Peter Schuster and Jörg Swetina. Consider a fitness landscape that contains a high but narrow

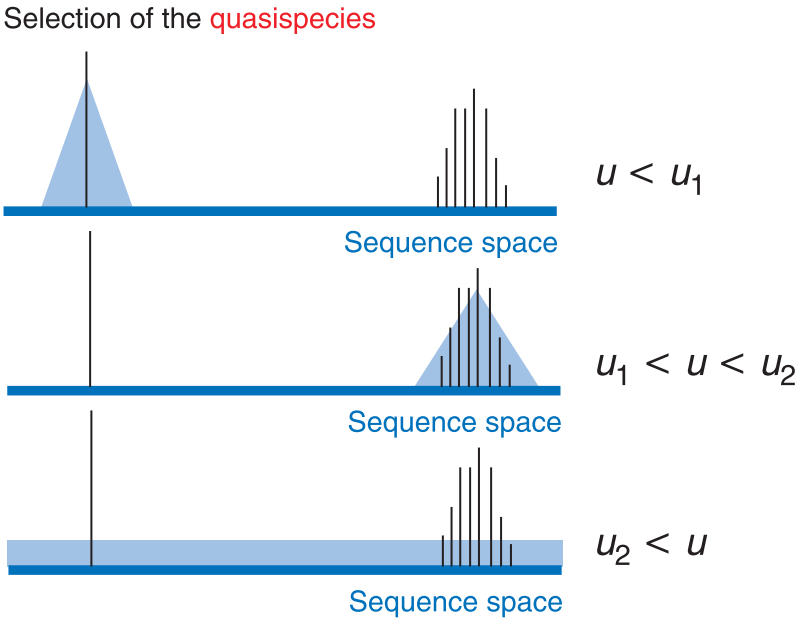

Figure 3.7 Consider a fitness landscape with two peaks. One is high but narrow, the other low but wide. If the mutation rate, u, is less than a critical value, u_1, then the higher peak is selected, indicated in blue. If the mutation rate, u, is greater than u_1, but less than the error threshold, u_2, then the lower peak is selected. If the mutation rate is greater than the error threshold, u_2, then neither peak can be maintained. For a given mutation rate, selection chooses the equilibrium quasispecies with maximum average fitness. "Survival of the fittest" is replaced by "survival of the quasispecies."

peak and in some distance a lower but broader peak (Figure 3.7). If the mutation rate is very small, the quasispecies at equilibrium will be centered around the higher peak. As the mutation rate increases, there is a sharp transition, and the quasispecies moves from the higher to the lower peak. The intuitive explanation is the following: for very small mutation rates only the maximum fitness matters, but for somewhat higher mutation rates the fitness of the neighboring sequences is also important. The second peak has a lower maximum fitness, but a broader ensemble of relatively good close-by neighbors. The first peak is like a brilliant person working alone, the second peak consists of a less brilliant person surrounded by a good team.

If mutation rates are sufficiently small, the quasispecies centered around the narrow peak has maximum fitness. But when mutation rates are higher, the quasispecies centered around the broader peak has maximum fitness. Beyond the error threshold neither peak can be maintained.

We conclude that selection does not always smile upon the fittest. For any given mutation rate, however, selection chooses the equilibrium distribution (the quasispecies) with maximum average fitness. "Selection of the fittest" is replaced by "selection of the quasispecies."

SUMMARY

- A quasispecies is a population of similar genomes.
- Quasispecies are formed by a mutation-selection process.
- In sequence space, all possible genomes of a certain length are arranged such that nearest neighbors differ by one point mutation. All sequences of length, L, can be arranged in a lattice that is embedded in an L-dimensional space.
- A fitness landscape is formed by assigning fitness values (reproductive rates) to all sequences. A fitness landscape is a high-dimensional mountain range over sequence space.
- Quasispecies live in sequence space and explore the fitness landscape.
- Quasispecies climb upward in the fitness landscape.
- The quasispecies equation describes deterministic evolutionary dynamics in terms of mutation and constant selection acting on an infinitely large population.
- Generically, the quasispecies equation has one globally stable equilibrium.
- At this equilibrium, the quasispecies consists not of solely the fittest genome but instead of a distribution of genomes in a mutation-selection balance.
- It is possible that this distribution does not contain the fittest genome at all. Hence "survival of the fittest" is replaced by "survival of the quasispecies."

- Adaptation is localization in sequence space. This is only possible if the mutation rate is below the error threshold.
- The error threshold states that the maximum possible mutation rate (per base) must be less than the inverse of the genome length (in bases).